Getting More Out of Large Databases and EHRs with Natural Language Processing and Artificial Intelligence

The Future Is Here

Bardia Khosravi, MD, MPH, MHPE, Pouria Rouzrokh, MD, MPH, MHPE, and Bradley J. Erickson, MD, PhD

Investigation performed at the Mayo Clinic, Rochester, Minnesota

Abstract: Electronic health records (EHRs) have created great opportunities to collect various information from clinical patient encounters. However, most EHR data are stored in unstructured form (e.g., clinical notes, surgical notes, and medication instructions), and researchers need data to be in computable form (structured) to extract meaningful relationships involving variables that can influence patient outcomes. Clinical natural language processing (NLP) is the field of extracting structured data from unstructured text documents in EHRs. Clinical text has several characteristics that mandate the use of special techniques to extract structured information from them compared with generic NLP methods. In this article, we define clinical NLP models, introduce different methods of information extraction from unstructured data using NLP, and describe the basic technical aspects of how deep learning-based NLP models work. We conclude by noting the challenges of working with clinical NLP models and summarizing the general steps needed to launch an NLP project.

Electronic health records (EHRs) have been widely adopted in hospitals and other care facilities to facilitate data recording, with the goals of helping patients and making health-care systems more accountable[1]. Some of the information stored in EHRs can be readily used for research purposes, including patients' vital signs and laboratory results and numerically encoded records of diagnoses and medical procedures. These are examples of structured data, i.e., data with a fixed format that can be directly acted on by a computer. Clinical researchers commonly use such structured data to enrich their understanding of factors influencing patients' outcomes.

Although structured data are an important part of EHRs, physicians record many aspects of care in an unstructured format, e.g., clinical visit notes, surgical notes, medication instructions, and radiology reports[2]. Such data rarely follow a specific format and are therefore not directly computable by most computer algorithms. While it is estimated that >80% of data recorded in EHRs are unstructured, extraction and processing of such data are tedious and may hinder clinical and research initiatives[3]. Therefore, there has been a tremendous effort to extract structured data from free-form text and use it to enrich existing research databases[4,5].

Artificial intelligence (AI) and specifically natural language processing (NLP) can help to extract specific data elements from unstructured text[6]. AI is a scientific field that explores how computers can do tasks requiring human intelligence. Machine learning (ML), a subcategory of AI, emphasizes computers learning the solution to a specific problem rather than having the solution hard-coded into them[7]. Deep learning (DL) is a category of ML that uses neural networks to learn the hidden patterns in the data. Finally, NLP is a discipline of AI that focuses on the automatic "understanding" of natural language[8], which we later discuss in more detail.

In this article, we first provide an overview of what clinical NLP is, how it differs from other NLP fields, and how it can help with information extraction in clinical research. We then examine technical aspects of how NLP models work and what steps may be taken to implement NLP research in an organization. Finally, we discuss some of the limitations and challenges in NLP that can hinder its use in clinical research.

Disclosure: This work was funded by a grant from the National Institute of Arthritis and Musculoskeletal and Skin Diseases (NIAMS) (P30AR76312) and the American Joint Replacement Research-Collaborative (AJRR-C) https://ajrrc.org/. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The **Disclosure of Potential Conflicts of Interest** forms are provided with the online version of the article (http://links.lww.com/JBJS/H134).

What Is Clinical NLP?

Clinical NLP is a subfield of NLP that helps computers understand clinical information from documents created by health-care professionals in clinical settings[9]. Clinical NLP must solve several challenges in processing free-form clinical texts that might not be considered critical in nonclinical contexts[10]. First, clinical texts are information-dense and communicate concepts between physicians with as few words as possible. Second, because of the high workload, there is a higher chance of spelling and grammatical errors and incomplete sentences during creation[11]. Finally, health-care professionals tend to use many abbreviations in their writing, which necessitates word disambiguation before extracting features, e.g., "MR" can refer to the English-language general word mister (Mr.), mental retardation, magnetic resonance, or mitral regurgitation, depending on the context[12].

Clinical NLP algorithms can be categorized into 2 general groups: rule-based and ML-based. The rule-based methods use predefined concepts to identify if a condition is met, or not met, in a sentence[13]. For example, Wyles et al. used rule-based NLP to detect implant cementation during total hip arthroplasty (THA) by checking whether the concept of cementation exists along with the concept of stem and shell in THA surgical reports[14]. Conversely, the ML-based methods, and more specifically DL-based ones, use many examples, typically with humans "highlighting" keywords to define the relationship, and hence are data-driven[15]. DL models develop semantic concepts of words and infer their relationship with each other on encountering similar words many times during their development or—as is commonly described in ML terminology—their training phase. As these models are trained on many examples, DL-based methods tend to be more robust with respect to the variability seen in human language patterns. While such powerful inference ability makes DL-based methods seem appealing and the first choice in deploying clinical NLP methods, this is not always the case. DL-based methods require a vast amount of data to form these interconnected webs of concepts, which makes training computers to use them difficult and expensive. However, transfer-learning can help with achieving good performance results with less data by "repurposing" an already trained model that solves a similar problem and fine-tuning it for the new problem.

Information Extraction with Clinical NLP

Clinical NLP can be used in several ways to help extract information from unstructured clinical notes and pour it into a structured format to be merged with already existing research databases. In this section, we briefly describe some of NLP's most important strategies for information extraction applied to clinical research.

Named Entity Recognition

Named entity recognition (NER) is the process of determining if a text, such as admission notes, contain an entity of interest (e.g., a particular procedure, diagnosis, or medication)[16]. For example, Karhade et al. used an ML-based NLP model to find patients who underwent reoperation after lumbar discectomy because of surgical site infection (their entity of interest) based on EHR notes[17]. Their model reached an area under the receiver operating characteristic curve (AUROC) of 0.99 in detecting this entity, whereas looking exclusively for the International Classification of Diseases (ICD) codes achieved an AUROC of 0.88.

Entity Linking

Entity linking (EL) is the task of associating a specific entity, like a diagnosis, with its corresponding value in a predetermined knowledge base, such as ICD codes[18]. EL pipelines usually start with an NER model, paired with a linking algorithm to find the association. For example, Koopman et al. used >400,000 death certificates to train an ensemble of ML-based NLP models to (1) determine if a cancer was the cause of death (NER task) and (2) pair the cancer type with its corresponding ICD code (EL task)[19]. They reached an F1 score of 0.94 for the NER task and an F1 score of 0.7 for the EL task. By utilizing transfer-learning, a more recent study reached an F1 score of 0.94 in this task[20].

Relation Extraction

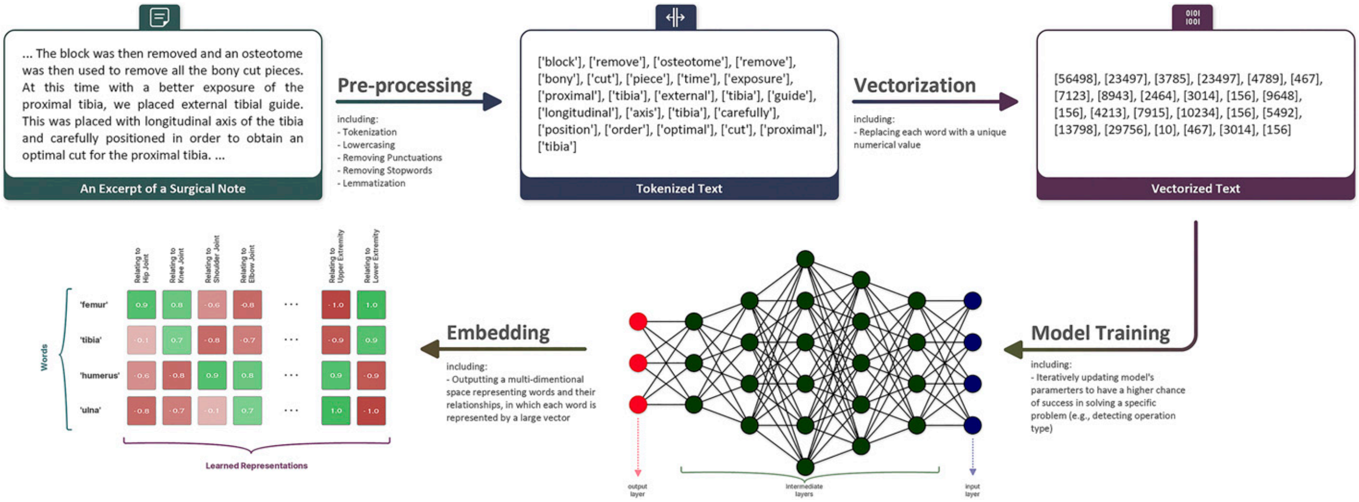
Another use of clinical NLP is relation extraction (RE). RE identifies the connection between different entities in a body of text[21], a critical step in understanding patients' medical records. For example, in the sentence "MRI shows avascular necrosis of the right femoral head and L4-L5 disc herniation," "MRI" is the procedure, connecting 2 medical conditions of "avascular necrosis of the right femoral head" and "L4-L5 disc herniation." Although both rule-based and DL-based NLP algorithms can undertake this task, DL-based models have shown superior performance, while also requiring much more data to train[22].

Medication Information Extraction

EHRs hold valuable information about patients' medication history. Most of these data are in the form of free text, which makes incorporating them in research databases impossible without first structuring them using an NLP algorithm. MedEx, the result of an early effort to identify medications in discharge notes, used rule-based search terms and in-text search strategies to create a structured history of patients' medications[23]. More recently, Mahajan et al. used a DL-based system to extract medication information, including name, administration frequency, route of administration, and dosage, to calculate the daily medication dose without using predefined rules[24]. These examples highlight the ability of NLP systems to reliably extract medication information from unstructured clinical notes and add them to previously created structured databases.

Behind the Scenes of an NLP Model

ML-based NLP models are more sophisticated and rely on automatic computations to figure out the necessary



An Excerpt of a Surgical Note

... The block was then removed and an osteotome was then used to remove all the bony cut pieces. At this time with a better exposure of the proximal tibia, we placed external tibial guide. This was placed with longitudinal axis of the tibia and carefully positioned in order to obtain an optimal cut for the proximal tibia. ...

Pre-processing

including:
- Tokenization
- Lowercasing
- Removing Punctuations
- Removing Stopwords
- Lemmatization

Tokenized Text

['block'], ['remove'], ['osteotome'], ['remove'], ['bony'], ['cut'], ['piece'], ['time'], ['exposure'], ['proximal'], ['tibia'], ['external'], ['tibia'], ['guide'], ['longitudinal'], ['axis'], ['tibia'], ['carefully'], ['position'], ['order'], ['optimal'], ['cut'], ['proximal'], ['tibia']

Vectorization

including:
- Replacing each word with a unique numerical value

Vectorized Text

[56498], [23497], [3785], [23497], [4789], [467], [7123], [8943], [2464], [3014], [156], [9648], [156], [4213], [7915], [10234], [156], [5492], [13798], [29756], [10], [467], [3014], [156]

Model Training

including:
- Iteratively updating model's paramerters to have a higher chance of success in solving a specific problem (e.g., detecting operation type)

Embedding

including:
- Outputting a multi-dimentional space representing words and their relationships, in which each word is represented by a large vector

Words	Relating to Hip Joint	Relating to Knee Joint	Relating to Shoulder Joint	Relating to Elbow Joint		Relating to Upper Extremity	Relating to Lower Extremity
'femur'	0.9	0.8	-0.6	-0.8	...	-1.0	1.0
'tibia'	-0.1	0.7	-0.8	-0.7	...	-0.9	0.9
'humerus'	-0.6	-0.8	0.9	0.8	...	0.9	-0.9
'ulna'	-0.8	-0.7	-0.1	0.7	...	1.0	-1.0

Learned Representations

Fig. 1
Overview of the steps of training in a DL-based NLP model.

rules for solving each task. DL-based models are also regarded as the state-of-the-art algorithms for solving many NLP tasks. Therefore, we focus on DL-based NLP algorithms in this section and use simple language to try to demystify how they work in 3 parts: data preprocessing, vectorization, and embedding (Fig. 1).

Data Preprocessing

Developing NLP algorithms requires the text to undergo several preprocessing steps before it can be used. Although the list of preprocessing options is very long, almost all NLP projects employ a few standard steps. First, each sentence in the given text is broken up into its constitutive words, a process called tokenization[25]. Next, all characters are converted to lowercase, and punctuation marks are removed[26]. All so-called stopwords (articles, prepositions, pronouns, and other linking words that do not change the meaning of a sentence) are also removed[27]. Then, a technique called lemmatization converts words to their base or dictionary form to reduce the diversity of repeated words[28]. While many more context-specific preprocessing steps are also available, their description is beyond the scope of this article. Interestingly, advances in DL-based models have made most of these steps less necessary compared with older ML-based or rule-based methods, as new DL-based models have the capacity to extract information from more word forms.

Vectorization of Input Words

DL models can be regarded as sophisticated mathematical equations with (1) a set of fixed parameters and (2) a set of input variables that represent text, imaging, and tabular or other data types. As in other mathematical operations, the parameters, inputs, and outputs of an equation are expected to be numbers. Therefore, we should convert our input words into numbers, a process called vectorization. When training an NLP model, these vectors are put into an equation with many parameters, and the computer tries to solve this equation by gradually identifying better values for its parameters; thus, the equation will have a higher chance of resulting in the expected output every time it sees the input data. The term *learning* in this context means that the model tweaks its parameters so that it reduces its error. This technique is termed *gradient descent,* the cornerstone of DL training in all ML applications, including, but not limited to, NLP[29].

Learning Word Embeddings

During the training, a model sees all input vectors in numerous text examples and develops an understanding of each word through learning word embeddings, which are high-dimensional vectors pertaining to the words of a text. For example, tibia and femur are likely to be close to each other in a sentence, while it would be less common for femur and humerus to appear in the same sentence or even paragraph. In this way, embeddings can encode the concept that certain words represent things that belong together. If certain gender-specific words like "he" or "she" are closely embedded with other terms like "king" or "queen," gender meanings can be inferred.

Last, we revisit what we meant by the concept that ML-based NLP models learn their own rules and need no hard-coding by humans. Rules that NLP models learn are a collection of well-crafted embedding vectors that they create and gradually optimize for different input words during the training. Embeddings of each input word contain many attributes of that word that the model considers important in solving the problem at hand. For example, the learned embedding vector for a given word may contain a numerical value denoting how important that word is in the context of fractures or may contain another value for denoting how the presence of that word should be weighed when answering a question about arthroplasty. This is where the magic of DL models happens.

Caveats and Biases

Clinical NLP has caveats that should be considered before starting NLP projects. In this section, we introduce such challenges and some of their widely used solutions.

Training Data

DL-based NLP algorithms require a considerable amount of data, on the order of millions of records, to learn robust embeddings. Transfer-learning has proven to be a viable option for training DL-based NLP models when there are not enough local data to train an NLP model from scratch[30]. Alternatively, rule-based NLP methods are already hard-coded, and therefore need no training data[31].

Patient Privacy

Privacy is among the core values of health-care systems and is regulated by international, national, and institutional entities. There are several approaches to protect patient privacy, including deidentification and pseudonymization. In deidentification, all patient identifiers (e.g., name and date of birth) are deleted, while pseudonymization replaces them with meaningful but not identifiable values to retain the correlation between different encounters[32]. Although helpful, neither of these approaches is perfect. Identifying data are sometimes placed in atypical locations (e.g., handwritten on images) or identity can be discerned because of rare medical conditions or extreme age, height, and/or weight. One way to protect privacy is to use federated learning. In federated learning, training is done on multi-institutional data, but no data leave the owner institution. Instead, each institution's data are analyzed locally and only the results of the analyses are aggregated from several centers to form a final model. Although this approach addresses important privacy issues of data sharing, it requires more computational power and complex infrastructures[33].

Amplifying Bias

There is a famous saying in the AI community that "a model is as good as the data it is trained on," resembling the popular saying "garbage in, garbage out." In this regard, existing biases in the training data will be amplified by the model. van Aken et al. trained several clinical NLP models using transfer-learning from different sources and found that the trained models were biased with respect to gender, ethnicity, and age[34]. For example, by letting a model know that a patient is of a specific race (compared with not mentioning their ethnicity), the model outputs a higher probability of drug abuse based on the same text describing another patient, although the data show that drug abuse is not different based on ethnicity[35]. Such differences are attributed to the source data used for pretraining or training the models. It is important that researchers be aware of such possible biases and try to address them during model development.

Explainability

Explainability means that a model should be able to provide insight on how it reached a decision. If these models are regarded as black-box solutions, clinicians and researchers may not trust them. Although not always applicable, several visualization techniques have been proposed to explain model output, including saliency highlighting or raw declarative representations[36]. While a discussion of such techniques is beyond the scope of this article, it is important that researchers apply them to their NLP model and check whether the model is paying attention to relevant sections of a text for specific outputs.

Implementing Clinical NLP

In this section, we propose a framework for performing an NLP project and note the resources it needs[37]. This framework consists of 4 major steps: (1) problem definition, (2) data acquisition and understanding, (3) model development, and (4) deployment.

The first step is to develop an understanding of the project and decide if the project requires NLP technology in the first place. If yes, then researchers should search the literature and see if this project has been already carried out before, and if so, what methods others used and if their code is available for rapid reproduction of their model.

The second step requires data acquisition and understanding. Researchers should decide what clinical notes to use and retrieve them from the EHR. As noted, it is best practice to also deidentify the data. Afterwards, an exploratory data analysis is done to understand the composition of the data, such as the number of records, demographic information of the patients, and categories of text (e.g., discharge summaries and notes to patients). If any data annotation is needed for further training of NLP models (e.g., marking the words that designate medication names and their dosages), it should also be done in this step.

In the third step, the NLP model is trained. The first branch point is whether to use a rule-based approach or, if there are enough data and ML expertise, to pursue an ML-based approach. As mentioned previously, ML and especially DL models rely on solving complex mathematical equations that need considerable computational power, such as strong graphics processing units (GPUs). One approach is to use cloud-based solutions that provide these resources online and charge on the basis of usage. After developing the model, its performance should be evaluated to make sure its findings are generalizable. For this, a portion of data are set aside and not used for training but only for testing the performance of the model.

Finally, the model is deployed and, ideally, its performance is evaluated continuously or periodically. This surveillance is important because of shifts in data composition and distribution that inevitably occur in the real world. For example, if there is a newly introduced medication, the model may fail to pick it up from the new clinical texts, requiring the researcher to retrain the model.

Conclusions

NLP can help with extracting information from unstructured text in EHRs to enrich existing research databases. Because of the properties of clinical free-form text, researchers face challenges when applying NLP algorithms to such text. There are different types of NLP models that can be utilized to extract unstructured data from clinical text, the selection of which can be based on data availability and task complexity. Although clinical NLP has reached promising results in terms of clinical information extraction, it still faces some challenges, such as needing large amounts of training data, protecting patient privacy, addressing underlying biases, and providing explainability. Such challenges need to be addressed before large-scale adoption of NLP models in research practice. ■

Bardia Khosravi, MD, MPH, MHPE[1,2]
Pouria Rouzrokh, MD, MPH, MHPE[1,2]
Bradley J. Erickson, MD, PhD[1]

[1]Radiology Informatics Lab (RIL), Department of Radiology, Mayo Clinic, Rochester, Minnesota

[2]Orthopedic Surgery Artificial Intelligence Laboratory (OSAIL), Department of Orthopedic Surgery, Mayo Clinic, Rochester, Minnesota

Email for corresponding author: AJRRC@mayo.edu

References

1. Kataria S, Ravindran V. Electronic health records: a critical appraisal of strengths and limitations. J R Coll Physicians Edinb. 2020 Sep;50(3):262-8.

2. Velupillai S, Suominen H, Liakata M, Roberts A, Shah AD, Morley K, Osborn D, Hayes J, Stewart R, Downs J, Chapman W, Dutta R. Using clinical Natural Language Processing for health outcomes research: Overview and actionable suggestions for future advances. J Biomed Inform. 2018 Dec;88:11-9.

3. Xiao C, Choi E, Sun J. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. J Am Med Inform Assoc. 2018 Oct 1;25(10):1419-28.

4. Tibbo ME, Wyles CC, Fu S, Sohn S, Lewallen DG, Berry DJ, Maradit Kremers H. Use of Natural Language Processing Tools to Identify and Classify Periprosthetic Femur Fractures. J Arthroplasty. 2019 Oct;34(10):2216-9.

5. Zeng J, Banerjee I, Henry AS, Wood DJ, Shachter RD, Gensheimer MF, Rubin DL. Natural Language Processing to Identify Cancer Treatments with Electronic Medical Records. JCO Clin Cancer Inform. 2021 Apr;5:379-93.

6. Li I, Pan J, Goldwasser J, Verma N, Wong WP, Nuzumlalı MY, et al. Neural Natural Language Processing for Unstructured Data in Electronic Health Records: a Review. arXiv. 2021. http://arxiv.org/abs/2107.02975

7. Bzdok D, Altman N, Krzywinski M. Statistics versus machine learning. Nat Methods. 2018 Apr;15(4):233-4.

8. Allen J. Natural Language Understanding. 2nd ed. Benjamin/Cummings; 1995.

9. Velupillai S, Mowery D, South BR, Kvist M, Dalianis H. Recent Advances in Clinical Natural Language Processing in Support of Semantic Analysis. Yearb Med Inform. 2015 Aug 13;10(1):183-93.

10. Tayefi M, Ngo P, Chomutare T, Dalianis H, Salvi E, Budrionis A, Godtliebsen F. Challenges and opportunities beyond structured data in analysis of electronic health records. Wiley Interdiscip Rev Comput Stat. 2021 Nov;13(6).

11. Dalianis H. Clinical Text Mining: Secondary Use of Electronic Patient Records. Springer; 2018.

12. Wang Y, Zheng K, Xu H, Mei Q. Interactive medical word sense disambiguation through informed learning. J Am Med Inform Assoc. 2018 Jul 1;25(7):800-8.

13. Fundel K, Küffner R, Zimmer R. RelEx—relation extraction using dependency parse trees. Bioinformatics. 2007 Feb 1;23(3):365-71.

14. Wyles CC, Tibbo ME, Fu S, Wang Y, Sohn S, Kremers WK, Berry DJ, Lewallen DG, Maradit-Kremers H. Use of Natural Language Processing Algorithms to Identify Common Data Elements in Operative Notes for Total Hip Arthroplasty. J Bone Joint Surg Am. 2019 Nov 6;101(21):1931-8.

15. Lopez MM, Kalita J. Deep Learning applied to NLParXiv. 2017. http://arxiv.org/abs/1703.03091

16. Meystre SM, Savova GK, Kipper-Schuler KC, Hurdle JF. Extracting information from textual documents in the electronic health record: a review of recent research. Yearb Med Inform. 2008:128-44.

17. Karhade AV, Bongers MER, Groot OQ, Cha TD, Doorly TP, Fogel HA, Hershman SH, Tobert DG, Schoenfeld AJ, Kang JD, Harris MB, Bono CM, Schwab JH. Can natural language processing provide accurate, automated reporting of wound infection requiring reoperation after lumbar discectomy? Spine J. 2020 Oct;20(10):1602-9.

18. Jin M, Bahadori MT, Colak A, Bhatia P, Celikkaya B, Bhakta R, Senthivel S, Khalilia M, Navarro D, Zhang B, Doman T, Ravi A, Liger M, Kass-hout T. Improving Hospital Mortality Prediction with Medical Named Entities and Multimodal Learning. arXiv. 2018. http://arxiv.org/abs/1811.12276

19. Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Automatic ICD-10 classification of cancers from free-text death certificates. Int J Med Inform. 2015 Nov;84(11):956-65.

20. Vashishth S, Joshi R, Dutt R, Newman-Griffis D, Rose C. Medtype: Improving medical entity linking with semantic type prediction. arXiv. 2020. https://arxiv.org/abs/2005.00460

21. Rink B, Harabagiu S, Roberts K. Automatic extraction of relations between medical concepts in clinical texts. J Am Med Inform Assoc. 2011 Sep-Oct;18(5):594-600.

22. Wei Q, Ji Z, Si Y, Du J, Wang J, Tiryaki F, Wu S, Tao C, Roberts K, Xu H. Relation Extraction from Clinical Narratives Using Pre-trained Language Models. AMIA Annu Symp Proc. 2020 Mar 4;2019:1236-45.

23. Xu H, Stenner SP, Doan S, Johnson KB, Waitman LR, Denny JC. MedEx: a medication information extraction system for clinical narratives. J Am Med Inform Assoc. 2010 Jan-Feb;17(1):19-24.

24. Mahajan D, Liang JJ, Tsou CH. Extracting Daily Dosage from Medication Instructions in EHRs: An Automated Approach and Lessons Learned. arXiv. 2020. http://arxiv.org/abs/2005.10899

25. Vijayarani S, Ilamathi MJ, Nithya M. Preprocessing techniques for text mining-an overview. Int J Computer Sci Comm Networks. 2015;5(1):7-16.

26. García S, Luengo J, Herrera F. Data Preprocessing in Data Mining. Springer; 2014.

27. Porter MF. An algorithm for suffix stripping. Program: electronic library and information systems. 1980 Jan 1;14(3):130-7

28. Plisson J, Lavrac N, Mladenic D. A rule based approach to word lemmatization. Proceedings of IS. 2004 May;3:83-6.

29. Andrychowicz M, Denil M, Gomez S, Hoffman MW, Pfau D, Schaul T, Shillingford B, de Freitas N. Learning to learn by gradient descent by gradient descent. In: Advances in neural information processing systems 29, 30th Annual Conference on Neural Information Processing Systems; 2016 Dec 5-10. Neural Information Processing Systems; 2016. p 3981-9.

30. Ruder S, Peters ME, Swayamdipta S, Wolf T. Transfer Learning in Natural Language Processing. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Tutorials; 2019 Jun 2. Association for Computational Linguistics; 2019. p 15-8.

31. Sagheb E, Ramazanian T, Tafti AP, Fu S, Kremers WK, Berry DJ, Lewallen DG, Sohn S, Maradit Kremers H. Use of Natural Language Processing Algorithms to Identify Common Data Elements in Operative Notes for Knee Arthroplasty. J Arthroplasty. 2021 Mar;36(3):922-6.

32. Giacomelli I, Jha S, Kleiman R, Page D, Yoon K. Privacy-Preserving Collaborative Prediction using Random Forests. AMIA Jt Summits Transl Sci Proc. 2019 May 6;2019:248-57.

33. Ma J, Zhang Q, Lou J, Ho JC, Xiong L, Jiang X. Privacy-Preserving Tensor Factorization for Collaborative Health Data Analysis. Proc ACM Int Conf Inf Knowl Manag. 2019 Nov;2019:1291-300.

34. van Aken B, Herrmann S, Löser A. What Do You See in this Patient? Behavioral Testing of Clinical NLP Models. arXiv. 2021. http://arxiv.org/abs/2111.15512

35. U.S. Department of Health and Human Services, Substance Abuse and Mental Health Services Administration. 2018 National Survey on Drug Use and Health. 2018. Accessed 2022 Feb 25. https://www.samhsa.gov/data/release/2018-national-survey-drug-use-and-health-nsduh-releases

36. Danilevsky M, Qian K, Aharonov R, Katsis Y, Kawas B, Sen P. A Survey of the State of Explainable AI for Natural Language Processing. arXiv. 2020. http://arxiv.org/abs/2010.00711

37. What is the team data science process? 2022. Accessed 2022 June 7. https://docs.microsoft.com/en-us/azure/architecture/data-science-process/overview